Registered Report

Advances in Methods and Practices in Psychological Science
April-June 2021, Vol. 4, No. 2, pp. 1–21

DOI: 10.1177/25152459211018199
www.psychologicalscience.org/AMPPS

How Do We Choose Our Giants? Perceptions of Replicability in Psychological Science

Manikya Alister, Raine Vickers-Jones, David K. Sewell, and Timothy Ballard
School of Psychology, The University of Queensland, Brisbane, AU-QLD, Queensland, Australia

Abstract

Judgments regarding replicability are vital to scientific progress. The metaphor of "standing on the shoulders of giants" encapsulates the notion that progress is made when new discoveries build on previous findings. Yet attempts to build on findings that are not replicable could mean a great deal of time, effort, and money wasted. In light of the recent "crisis of confidence" in psychological science, the ability to accurately judge the replicability of findings may be more important than ever. In this Registered Report, we examine the factors that influence psychological scientists' confidence in the replicability of findings. We recruited corresponding authors of articles published in psychology journals between 2014 and 2018 to complete a brief survey in which they were asked to consider 76 specific study attributes that might bear on the replicability of a finding (e.g., preregistration, sample size, statistical methods). Participants were asked to rate the extent to which information regarding each attribute increased or decreased their confidence in the finding being replicated. We examined the extent to which each research attribute influenced average confidence in replicability. We found evidence for six reasonably distinct underlying factors that influenced these judgments and individual differences in the degree to which people's judgments were influenced by these factors. The conclusions reveal how certain research practices affect other researchers' perceptions of robustness. We hope our findings will help encourage the use of practices that promote replicability and, by extension, the cumulative progress of psychological science.

Keywords

replicability, reproducibility, metascience, attitudes, open data, open materials, preregistered

Received 6/25/19; Revision accepted 4/20/21

Many people have argued that psychology is experiencing a "crisis of confidence" given several prominent failures to replicate seemingly established findings (Earp & Trafimow, 2015; Pashler & Wagenmakers, 2012). Concerns about replicability have led to several highly publicized, large-scale replication attempts. The results of these studies have not been especially encouraging; replicability rates have ranged from 39% to 77% using a criterion of $p < .05$ (Camerer et al., 2018; Klein et al., 2014; Open Science Collaboration, 2015). Among the most well known was the attempt by the Open Science Collaboration (2015) to replicate 100 experiments reported in high-ranking psychology journals; only 39% of the original effects were replicated. A similar attempt to replicate 21 significant results from social science articles published in the journals *Science* and *Nature* successfully replicated only 62% of effects (Camerer et al., 2018). Three Many Labs replication attempts have been conducted in which scientists from different labs performed the same experiments to gain large, heterogeneous samples to test the effects found in prior research. In the first Many Labs project, 77% of original findings—about 13 different effects—were replicated (Klein et al., 2014). In the second project, which examined 28 different effects, only 54% were replicated (Klein et al., 2018). The third project examined 10 effects, and 40% were replicated (Ebersole et al., 2016).

Corresponding Author:
Timothy Ballard, School of Psychology, The University of Queensland, Brisbane, AU-QLD, Queensland, Australia
E-mail: t.ballard@uq.edu.au

Although any individual failure to replicate could be due to a multitude of reasons—and not necessarily indicative of a larger systemic problem—mounting numbers of replication failures are difficult to dismiss. The relatively low replication rates reported by the studies mentioned above have been held up as prima facie evidence of a significant discipline-wide problem. To address this problem, there have been a number of proposals to change the way research is conducted. Examples of such practices include preregistration (Jonas & Cesario, 2016; Nosek et al., 2018; Wagenmakers et al., 2012), making data publicly available (Asendorpf et al., 2013; Miguel et al., 2014), and increasing sample sizes (Cohen, 1962; Tversky & Kahneman, 1971). Other researchers have advocated for different statistical procedures, such as using Bayes's factors instead of null-hypothesis significance testing (Etz & Vandekerckhove, 2016) or changing the default cutoff p value from .05 to .005 for new discoveries (Benjamin et al., 2018). What is not currently known is the extent to which such practices influence perceptions of replicability. Put simply, does adherence to such practices affect other researchers' confidence in the replicability of a result?

Judgments regarding replicability are vital to the cumulative progress of science. The accumulation of scientific knowledge results from an ongoing process whereby new discoveries build on previous discoveries. The nature of this process is summarized elegantly by the idea of "standing on the shoulders of giants." According to this metaphor, scientists can see further than their predecessors only because they stand atop a body of knowledge those predecessors have built. For this process to be successful, however, it is important that scientists can accurately judge the robustness of the existing knowledge on which they build. Attempts to build on findings that turn out to be false could mean a great deal of time, effort, and money wasted. It is therefore in researchers' best interest to choose their giants carefully.

Although it can be difficult to fully articulate what signals whether a finding is robust or, more generally, the "quality" of a study, it appears that researchers are quite capable of judging the replicability of published results. In 2012 and 2014, prediction markets were run to ascertain the ability of psychologists to predict whether studies included in the Reproducibility Project would be successfully replicated (Dreber et al., 2015). The prediction market closed with a market price above 50 for 29 of the 41 replications attempted, therefore correctly anticipating the outcomes for 71% of the sample. Conversely, Camerer et al. (2016) ran a prediction market for replications of 18 experimental economics studies and did not find a significant relationship between the market and the actual replication rate. However, Camerer et al. (2018) later ran a prediction market for replications of 21 studies published in *Nature* and *Science* and found that the market beliefs were highly correlated with successful replication. A fourth prediction market run to assess the replication likelihood of the 24 studies included in Many Labs 2 (Klein et al., 2018) found that the market correctly predicted the results of 75% of the replications (Forsell et al., 2019). The overall success of these prediction markets suggests that researchers are sensitive to the features of research studies that influence the replicability of findings. An outstanding question, however, is what those features are.

Anecdotally, the quality of research has been inferred, at least in part, by the prestige of the journals in which it is published. Likewise, the reputations of individual researchers have sometimes been used to informally assess the quality of published research. It is not known, however, whether these traditional indicators of quality are also used to judge the replicability of findings. More recently, it would seem that the research practices intended to improve replicability described above would constitute salient criteria on which to judge replicability. However, very little research has been done to establish which, if any, of these newer practices actually influence confidence in the replicability of a finding.

The aim of this study was to examine how various research practices or features influence other researchers' confidence in the replicability of a finding. We did this by surveying corresponding authors of articles published in psychology journals between 2014 and 2018. Our analysis presented here focuses on three research questions. The first question concerns the extent to which each feature influences average confidence in replicability. For example, on average, how strongly are people influenced by features such as sample size, the use of preregistration, or the use of a particular statistical method? The second research question concerns the underlying factors that influence judgments of replicability. For example, are there distinct themes that people consider when making these evaluations? The third question concerns whether there are different profiles of beliefs regarding the effectiveness of certain research practices for fostering replicability. For example, are certain people's judgments of replicability more strongly influenced by issues relating to statistical methods but other people's judgments more strongly influenced by the presence of preregistered hypotheses and publicly available data? By clarifying understanding of what it is that researchers perceive as being relevant to or signaling the likelihood of replicability, we hoped to identify research practices that are likely to be widely adopted as well as those that might require further justification. This study also provides benchmark data to compare perceptions of what signals replicability with research

practices that actively increase replicability. Determining how attuned psychology is, as a discipline, to the factors that promote replicability is essential for efficiently and effectively increasing the rigor with which psychological research is conducted.

Disclosures

Preregistration

The preregistration documentation for this study can be found at https://osf.io/9nwrd.

Data, materials, and online resources

All the materials, code, and de-identified data from this study can be found at https://osf.io/dj2fx/.

Reporting

We report how we determined our sample size, all data exclusions, and all measures in the study.

Ethical approval

The protocol was approved by the University of Queensland Faculty of Health and Behavioural Science Human Ethics Committee (Protocol 2020000887). The study was carried out in accordance with the provisions of the Declaration of Helsinki.

Method

Participants

Our sample comprised psychological scientists who were recruited using the procedure described by Field et al. (2018; for full details of their procedure, see https://osf.io/s7a3d/). Following their procedure, we searched for journal articles published between 2014 and 2018 that appear under the following categories in the Web of Science database: "Psychology Multidisciplinary," "Psychology Applied," "Psychology Clinical," "Psychology Social," "Psychology Educational," "Psychology Experimental," "Psychology Developmental," "Behavioral Sciences," and "Psychology Mathematical." This search yielded 14,251 articles. We extracted the e-mail address of the corresponding author of each article. After we removed duplicates, 9,017 unique e-mail addresses remained. Our full data-extraction procedure can be found on OSF (https://osf.io/jqyux/). For this study, we used 5,000 of these unique e-mail addresses and reserved the rest for a potential follow-up study. We expected a response rate similar to that of Field et al., which was approximately 12.5%. This response rate would result in a sample size of approximately 625 participants.

We compared our demographic variables with data from the National Science Foundation's (NSF; 2017) Survey of Doctoral Recipients to assess the representativeness of our sample. The NSF data contain demographic details of more than 100,000 individuals with a research doctoral degree in psychology from an institution in the United States. Our sample was generally well aligned with the NSF sample, although our sample contained a higher proportion of males, and participants in our sample tended to be slightly younger. In the NSF sample, 41% of these individuals identified as male, and 59% identified as female. Our sample contained a slightly lower percentage of females (61% male, 35% female, 1% preferred to self-describe, 3% preferred not to say). In the NSF sample, 7% of individuals with a U.S. doctoral degree in psychology were under the age of 35 (19% in our sample), 10% were 35 to 39 years old (17% in our sample), 11% were 40 to 44 years old (15% in our sample), 12% were 45 to 49 years old (6% in our sample), 11% were 50 to 54 years old (10% in our sample), 12% were 55 to 59 years old (6% in our sample), 13% were 60 to 64 years old (4% in our sample), and 23% were 65 to 75 years old (9% in our sample). Likewise, our sample contained participants who had more recently received their PhD. In the NSF sample, the PhD was awarded 5 or fewer years prior for 12% of respondents (22% in our sample), 6 to 10 years prior for 14% of respondents (18% in our sample), 11 to 15 years prior for 13% of respondents (15% in our sample), 16 to 20 years prior for 14% of respondents (9% in our sample), 21 to 25 years prior for 13% of respondents (8% in our sample), and more than 25 years prior for 34% of respondents (16% in our sample).

Materials and Procedure

The survey was created using Qualtrics and was distributed using a shareable link sent by e-mail. Reminder e-mails were sent 1 and 3 weeks after the initial send date. Data collection concluded 1 month after the initial survey was sent out. The full survey can be accessed at https://osf.io/2dxe6/. The first survey item asked participants to estimate the percentage of randomly selected studies from the psychological literature that would be successfully replicated if a high-powered replication attempt (with a large enough sample size to measure the effect precisely) using the exact same methods and statistical analyses were to be conducted. Responses were provided on a scale from 0% to 100%.

In the following section, participants were asked to rate how much specific study attributes (e.g., research practices) would increase or decrease their confidence that a randomly selected effect from the psychological literature would be replicated. Participants were asked to consider 76 specific study attributes, each requiring

an independent confidence rating. For each participant, the attributes were presented in random order, and responses were provided on a Likert scale ranging from −5 (*substantial decrease in confidence*) to +5 (*substantial increase in confidence*); 0 indicated *no change in confidence*. The set of study attributes presented to participants spanned a wide range of issues that might bear on the replicability of a finding. Examples include "The original study was pre-registered," "The conclusions of the original study were based on analyses planned before data collection," "The conclusions of the original study were based on post-hoc analyses," and "The original study was from the field of social psychology." See Appendix A for the full set of questions.

Some of these items address factors that have been considered in large-scale replication studies assessing the objective replicability of studies (e.g., sample size, experience of the research team, surprisingness of the effect). Others address attributes that have not been the focus of previous research but nevertheless may bear on perceptions of replicability (e.g., the type of journal in which the original study was published, the length of the article in which the original study was reported). We also included items relating to the subfield of psychology from which the study originated because we were interested in examining whether there was variance in the perceived replicability of research in different subfields. Additional items were based on feedback from an outside team of researchers who have collected data on experts' predictions of the replicability of research and the reasons considered during these assessments. Fifteen members of the repliCATS team (https://replicats.research.unimelb.edu.au/) responded to e-mails from one of our authors inviting them to provide a list of features they have observed researchers considering when assessing the replicability of research. The repliCATS project has been running since April 2019 as a component of the U.S. government's Systematizing Confidence in Open Research and Evidence program. Finally, we included control items that presented information that could not plausibly influence replicability (e.g., "The lead author on the original study was right-handed").

After responding to the 76 attributes, participants were presented with a set of items that assessed demographic details such as their age, gender ("male," "female," "prefer not to say," or "prefer to self-describe"), and career stage ("undergraduate student," "postgraduate student," "early career academic," "mid-career academic," "senior academic," "left academia after finishing PhD," "left academia before finishing PhD," or "none of these apply to me"). Participants who described themselves as an early-career academic, a midcareer academic, a senior academic, or as having left academia after finishing their PhD were also asked to report the number of years it had been since they were awarded their PhD. Participants were also asked which field of psychology they were most interested in or most associated with, depending on their career phase ("clinical," "cognitive," "developmental," "industrial/organizational," "evolutionary," "neuropsychology," "social," "quantitative," "biological," "health," "human factors," "unsure," "other," or "prefer not to say"). Participants who had left academia or were postgraduate students, early-career academics, midcareer academics, or senior academics were asked which field they associated with, whereas participants who were undergraduates or who selected the "none of these apply to me" response were asked which field they were most interested in.

The next set of items assessed participants' knowledge of and familiarity with issues surrounding replicability in psychology. Participants were asked to rate how familiar they were with the current debates surrounding the replicability of psychological science (scale from 0 = *not at all familiar* to 10 = *extremely familiar*), how important they considered the issue of replicability in psychological science to be (scale from 0 = *not at all important* to 10 = *extremely important*), and the extent to which they believed psychological science is currently experiencing a replication crisis (scale from 0 = *I do not believe this at all* to 10 = *I believe this very strongly*). Participants who identified as postgraduate students or academics were also asked how often before data collection they (a) preregistered their studies, (b) made a priori hypotheses, and (c) ran formal power analyses. Each of the questions in this set had an option to not respond.

The final set of items pertained to the participants' publication records. Participants were asked (a) what their h-index was, (b) how many publications they had, and (c) how many citations they had. Unfortunately, as a result of issues with potential identifiability, we could not release these data publicly. However, the raw publication data will be released on request to individual researchers who are committed to maintaining confidentiality. The survey took approximately 12 min to complete.

Results

Seventy-six of the 5,000 e-mails were undeliverable (e.g., recipient moved institutions, was on leave, or had retired). Therefore, the total number of e-mails successfully sent was 4,924. Two-hundred ninety-nine surveys were incomplete and consequently removed from data analysis. The survey was completed by 503 participants, which equated to a response rate of 10% (which compares favorably with that of Field et al., 2018, who had a response rate of approximately 9.5% for completed data).

Demographic and background information

Of the 503 respondents, 306 were male, 176 were female, 16 preferred not to say, and five preferred to self-describe. The average age of the sample was 45.9 years (59 did not provide their age). The sample contained one undergraduate student, 24 postgraduate students, 130 early-career academics, 140 midcareer academics, 162 senior academics, 19 individuals who left academia after finishing their PhD, one who left academia before finishing their PhD, 19 who said that none of these options applied to them, and seven who preferred not to respond. On average, participants who identified as academics or who had left academia after completing their PhD finished their PhD 15.9 years before this study. The percentages of participants who most strongly associated with (or were most interested in) each field of psychology were as follows: cognitive psychology, 17%; clinical psychology, 14%; developmental psychology, 7%; industrial/organizational psychology, 9%; evolutionary psychology, 5%; neuroscience/neuropsychology, 2%; social psychology, 24%; health psychology, 7%; quantitative psychology, 6%; human factors psychology, 2%; biological psychology, 3%; unsure, 1%; and other, 19%. Participants who selected the "other" option listed subfields such as counseling psychology, educational/school psychology, history of psychology, personality psychology, and sport psychology. Participants were able to select more than one answer, so the percentages sum to more than 100%.

On average, participants rated their familiarity with the current debates surrounding the replicability of psychological science to be 6.8 out of 10 (SD = 2.2; 1% preferred not to respond). The average degree of importance participants placed on the issue of replicability in psychological science was 8.3 out of 10 (SD = 1.7; 1% preferred not to respond). The average strength of belief that psychological science is currently experiencing a replication crisis was 6.7 out of 10 (SD = 2.1; 2% preferred not to respond). On average, the self-reported frequencies of preregistration, making a priori hypotheses, and running a formal power analysis were, respectively, 3.8 (SD = 3.6; 12% preferred not to respond; 10% indicated that this does not apply to their research), 8.4 (SD = 1.9; 11% preferred not to respond; 6% indicated that this does not apply to their research), and 6.4 (SD = 3.1; 11% preferred not to respond; 7% indicated that this does not apply to their research) out of 10.

Overview of analyses

We analyzed the results in three parts. The goal of Part 1 was to get an overall sense of the influence that each attribute has on confidence in replicability. In this part, we compared the distributions of confidence ratings across the 76 items. The goal of Parts 2 and 3 was to identify different profiles of beliefs regarding the factors that signal replicability. In Part 2, we conducted an exploratory factor analysis to identify the latent themes surrounding the factors that influence people's confidence in replicability. In Part 3, we conducted a clustering analysis on the factor scores obtained in Part 2 to examine whether there were distinct subgroups of participants whose judgments of replicability were influenced by different issues (or in different ways).

Research Question 1: How Does Each Feature Influence Average Confidence in Replicability?

On average, participants believed that 53% (SD = 18.57%) of studies would successfully replicate if a random sample of studies were drawn from the psychological literature and subjected to a high-powered replication attempt using the exact same methods and statistical analyses. The relatively low expected rate of replication underscores the need to identify specific study factors that increase, or decrease, confidence in research that replicates findings.

For the distributions of the ratings for each item, see Figures 1 and 2. Note that the mean ratings ranged from −2.57 ("original study had low statistical power") to 2.95 ("original result has been successfully replicated using same methods"). Information that increased confidence in the replicability of the original study included the existence of previous replications, high power or sample size, open data and materials, robustness of the phenomenon across contexts, and a representative sample of participants; analyses that were preregistered, planned in advance, or the original study was a Registered Report also increased confidence. Information that decreased confidence in the replicability of the original study included the study having low power or a small sample size, data and methods that are not openly available, a phenomenon that is highly context dependent, or results that involve a three-way interaction, are based on p values just below the threshold for significance, are surprising, or are based on post hoc analyses. There were also several items that did not have a strong influence on confidence in replicability, such as the results having practical applications, the career stage of the researcher, or, as expected, whether the researcher was left- or right-handed.

Research Question 2: What Are the Factors That Influence Judgments of Replicability, and What Are Their Constituent Features?

To answer this question, we conducted a Bayesian exploratory factor analysis on participants' responses to

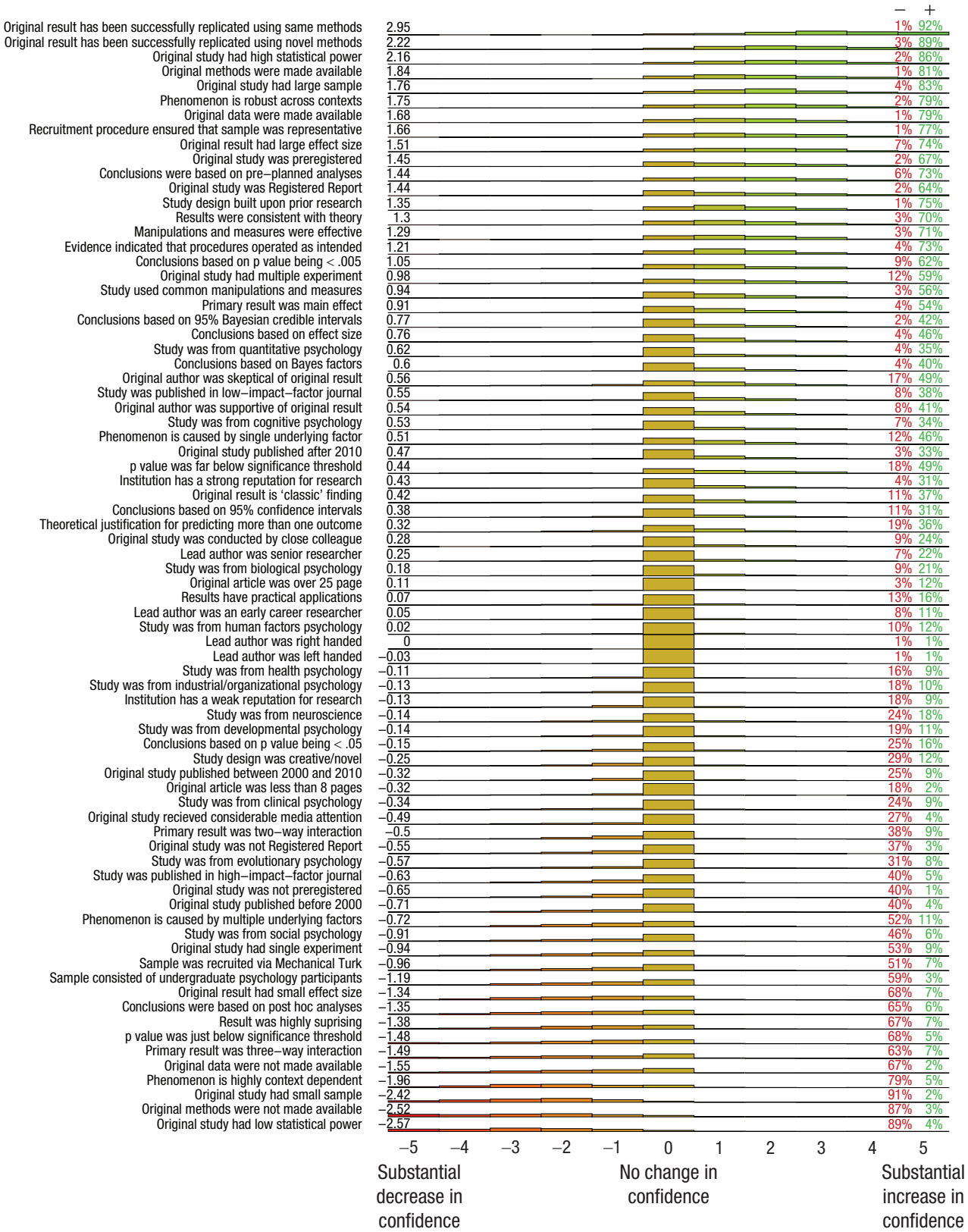

Fig. 1. The distribution of ratings for all confidence items. The height of each rectangle corresponds to the relative frequency of that rating being chosen. The color of each rectangle indicates the degree to which the presence of the attribute in question increased or decreased confidence in replicability, as indicated by the color key. The values shown on the left side of the plot indicate the mean rating for each attribute. The values on the right side of the plot indicate the percentage of participants for whom each respective factor decreased confidence (red) and increased confidence (green). For the full color figure, see the online version of the article. A table containing the means and standard deviations for all of the confidence items can be found at https://osf.io/b2wmj/.

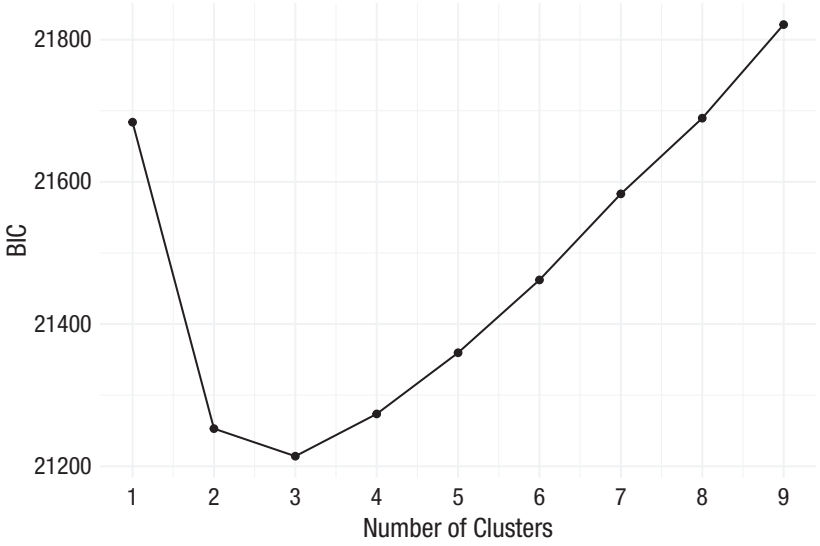

Fig. 2. Results of the model selection: the Bayesian information criterion (BIC) values associated with alternative Gaussian mixture models instantiating different numbers of clusters. Lower BICs indicate a better trade-off between fit and parsimony.

each item. The goal of this analysis was to (a) determine the number of underlying factors that best characterized participants' responses and (b) understand the patterns of relationships between specific features or practices and the underlying factors. The analysis was carried out using the Bayesian approach to exploratory factor analysis developed by Conti et al. (2014), which is implemented by the *BayesFM* package (Piatek, 2019) in R. The benefit of this approach is that it allows the researcher to simultaneously infer the number of factors and estimate the factor loadings rather than requiring an initial decision about the number of factors to retain before item loadings can be estimated. It also takes advantage of Bayesian methods for model selection by computing the posterior model probability for each possible number of factors, which provides information regarding the relative evidence for the different factor structures.

This approach uses Markov chain Monte Carlo (MCMC) methods to implement a factor model in which each item is allowed to load onto no more than one latent factor. The analysis estimates the number of factors as well as the loading matrix and factor correlations under each factor structure considered. Conti et al. (2014) demonstrated that this method accurately recovers the true structure of the factor-loading matrix in the vast majority of cases when the sample size is at least 500. This suggests that our sample size should be sufficiently large to enable reliable inferences to be made regarding the underlying structure of the items.

Following Conti et al. (2014), we constrained the model such that each factor was required to have at least three items that loaded only onto that factor. We also retained their default priors (see Appendix B). The analysis requires the user to specify the maximum number of factors permitted. We set this maximum to six. We believed this setting would strike a good balance by allowing considerable variety in the factor structures that could emerge without allowing so many factors that the underlying themes would be difficult to interpret. These settings resulted in a prior on the number of factors for which approximately 8% of the mass was allocated to the one-factor model, 30% was allocated to the two-factor model, 38% was allocated to the three-factor model, 20% was allocated to the four-factor model, 4% was allocated to the five-factor model, and less than 1% was allocated to the six-factor model.

The MCMC analysis included four chains (with unique, randomly generated starting values). Each chain had a burn-in period of 20,000 samples. After burning in, each chain produced 20,000 more samples. Therefore, the final analysis was based on 80,000 samples (i.e., 4 Chains × 20,000 Samples Per Chain). We assessed convergence by computing the Gelman-Rubin (R) statistic (Gelman & Rubin, 1992); values lower than 1.1 indicated successful convergence. We also inspected the trace plot of each parameter for evidence of mixing and stationarity. The chains demonstrated good convergence, with very little autocorrelation in the final samples. This suggests that the approximated posterior distributions are likely to be highly representative of their underlying distributions (Kruschke, 2015).

We compared the evidence for the different factor structures by examining the posterior probability for each number of factors. The analysis revealed that a structure containing six distinct factors was most probable, with 100% of the posterior mass favoring this structure.

We interpret the item loadings and factor correlations from the six-factor model given that this model was deemed most probable. Table 1 shows the loadings of all the items onto the factors, which are ordered from the most to least variance explained. Factor 1 contains items that signal that the original study may have suffered from weak methodology (e.g., low sample size/power, the use of a convenience sample, lack of open data/methods). Factor 2 contains items that might be viewed as suggestive of questionable research practices[1] such as HARKing or *p*-hacking (e.g., results that are surprising, involve complex interactions, are just below the threshold for significance, or are based on analyses that were not planned in advance). This factor also contained the highest number of subdiscipline-related items; the social psychology item loaded the highest onto this factor, followed by the evolutionary psychology and clinical psychology items.

Factor 3 contained items suggestive of a rigorous analysis (e.g., a large sample, high power, an analysis that was planned in advance). The subdisciplines of cognitive and quantitative psychology also loaded onto this factor. Factor 4 contains items relating more directly to the ease of conducting a replication study (e.g., the existence of previous replications, the presence of open methods and data, and the effectiveness of the methodology). Factor 5 contains items that reflect the robustness of the conclusions at a higher level (e.g., consistency with theory and prior research, robustness across contexts or experiments). Items relating to the use of Bayesian statistical methods also loaded onto this factor. Factor 6, which explained the least amount of variance, was the most difficult factor to identify a coherent theme for. We refer to this factor as the Established Convention factor because many of the items might be viewed as traditional indicators of replicability (e.g., the status of the researcher or institution, the finding being regarded as "classic," and the use of standard statistical tools such as effect size and confidence intervals). There was also one item that did not load onto any factor (the lead author having been left-handed).

As can be seen in Table 2, many of the factors are correlated. Factors 1 and 2 are positively correlated, which makes sense given that both of these factors relate negatively to replicability. Factors 3 through 6 are all positively correlated with each other and negatively correlated with Factors 1 and 2. As can be seen in Table 2, several of the factors correlate fairly strongly with one another, which suggests that some of the factors tap into similar constructs. For example, both Rigorous Analysis and Ease of Replication incorporate methodological elements that arguably serve similar purposes (e.g., effective research design [Ease of Replication] and large sample size [Rigorous Analysis]). It is possible that these factors are conceptually separable but are correlated because they are often perceived to co-occur.

To examine the robustness of our results to changes in the priors, we conducted a prior sensitivity analysis. To do this, we ran a follow-up analysis with different prior values for the concentration parameter that influences the mass allocated to the different numbers of factors. The value of this parameter in the analysis reported above was .17 (which is the recommended setting for a model with a maximum of six factors). We ran follow-up analyses with this parameter set to .05 (a setting that more strongly favors solutions with fewer numbers of factors) and .5 (which favors solutions with more factors). Both alternative settings resulted in a six-factor structure—the same number of factors as the model reported above. When the concentration parameter was set to .05, the pattern of loadings was virtually identical to the model above. When the concentration parameter was set to .5, items from Factors 4 (Ease of Replication) and 5 (Robustness of Conclusions) and most of the items from Factor 3 (Rigorous Analysis) in the above model merged into a single factor that explained 71% of the variance captured by the model. Remaining items from Factor 3 loaded onto a second factor (19% of the variance), and most items from Factor 1 (Weak Methodology) in the above model loaded negatively onto this factor. The contents of the third and fourth factors in this model (7% and 2% of the variance, respectively) were comparable with Factors 2 (Potential for Questionable Research Practices [QRPs]) and 6 (Established Convention) in the above model. The fifth and sixth factors each explained less than 1% of the variance captured by the model. Most of the items on the fifth factor related to the subdiscipline, whereas most items on the sixth factor were remaining items from Factor 1 (Weak Methodology) in the above model.

These results suggest that although the choice of prior exerts some influence on the distribution of thematically related groups of items across factors, the item groupings themselves are fairly stable. In other words, items that are grouped together in one model tend to be grouped together in the other models. The two exceptions are items that loaded onto Factors 1 and 3 in the original model. Under an alternative prior, with the concentration parameter set to .5, items loaded onto these factors are each split across two different factors. However, these factors are highly correlated ($r = -.83$ for the Factor 1 items; $r = .90$ for the Factor 3 items), which suggests that they might best be viewed as a single factor anyway.

Table 1. Loading of Each Item Onto the Relevant Latent Factor From the Six Factor Model

	Factor					
Item	1	2	3	4	5	6
21. Original study had low statistical power	2.81 [2.61, 3.03]					
37. Original methods were not made available	2.71 [2.5, 2.93]					
50. Original study had small sample	2.68 [2.49, 2.88]					
64. Phenomenon is highly context dependent	2.17 [1.99, 2.36]					
35. Original data were not made available	1.71 [1.54, 1.88]					
73. Original result had small effect size	1.51 [1.36, 1.67]					
53. Sample consisted of undergraduate psychology participants	1.33 [1.18, 1.49]					
67. Original study had single experiment	1.10 [0.95, 1.24]					
54. Sample was recruited via Mechanical Turk	1.04 [0.89, 1.19]					
63. Phenomenon is caused by multiple underlying factors	0.88 [0.75, 1.02]					
52. Institution has a weak reputation for research	0.20 [0.12, 0.28]					
26. Primary result was three-way interaction		1.96 [1.80, 2.14]				
47. *p* value was just below the threshold for significance		1.92 [1.76, 2.09]				
19. Result was highly surprising		1.79 [1.63, 1.95]				
11. Conclusions were based on post hoc analyses		1.65 [1.49, 1.82]				
13. Study was from social psychology		1.32 [1.18, 1.47]				
25. Primary result was two-way interaction		0.92 [0.81, 1.04]				
31. Original study was published before 2000		0.89 [0.77, 1.01]				
2. Original study was not preregistered		0.84 [0.74, 0.95]				
68. Original study received considerable media attention		0.80 [0.69, 0.91]				
16. Study was from evolutionary psychology		0.77 [0.64, 0.91]				
4. Original study was not Registered Report		0.73 [0.62, 0.84]				
61. Study design was creative or novel		0.65 [0.54, 0.75]				
17. Study was from clinical psychology		0.57 [0.46, 0.68]				
32. Original study was published between 2000 and 2010		0.55 [0.46, 0.65]				
5. Conclusions were based on *p* values being < .05		0.51 [0.40, 0.62]				

(continued)

Table 1. *(continued)*

	Factor					
Item	1	2	3	4	5	6
14. Study was from neuroscience		0.48 [0.36, 0.61]				
43. Study was from health psychology		0.37 [0.28, 0.46]				
15. Study was from developmental psychology		0.34 [0.25, 0.44]				
18. Study was from industrial/ organizational psychology		0.30 [0.19, 0.41]				
22. Original study had high statistical power			2.39 [2.21, 2.57]			
49. Original study had large sample			1.97 [1.81, 2.14]			
3. Original study was Registered Report			1.77 [1.61, 1.99]			
10. Conclusions were based on preplanned analyses			1.75 [1.58, 1.92]			
1. Original study was preregistered			1.71 [1.55, 1.91]			
38. Conclusions were based on *p* values being < .005			1.26 [1.12, 1.42]			
24. Primary result was main effect			1.07 [0.94, 1.19]			
45. Study was from quantitative psychology			0.74 [0.63, 0.85]			
75. Replicating team was skeptical of original result			0.72 [0.58, 0.86]			
12. Study was from cognitive psychology			0.71 [0.61, 0.83]			
48. *p* value was far below the threshold for significance			0.66 [0.48, 0.84]			
69. Original study was conducted by close colleague			0.41 [0.31, 0.51]			
71. Original result has been successfully replicated using same methods				3.07 [2.86, 3.31]		
72. Original result has been successfully replicated in novel context				2.40 [2.22, 2.60]		
36. Original methods were made available				2.08 [1.92, 2.25]		
34. Original data were made available				1.90 [1.75, 2.06]		
55. Recruitment procedure ensured that sample was representative				1.87 [1.72, 2.03]		
58. Manipulations and measures were effective				1.47 [1.34, 1.62]		
30. Original article less than 8 pages				−0.33 [−0.42, −0.25]		
65. Phenomenon is robust across contexts					1.99 [1.84, 2.15]	
74. Original result had large effect size					1.74 [1.58, 1.91]	

(continued)

Table 1. *(continued)*

	Factor					
Item	1	2	3	4	5	6
60. Study design was built upon prior research					1.61 [1.49, 1.75]	
20. Results were consistent with theory					1.58 [1.46, 1.72]	
59. Evidence indicated that procedures and measures operated as intended					1.42 [1.30, 1.55]	
57. Study used common manipulations and measures					1.25 [1.13, 1.37]	
66. Original study had multiple experiments					1.23 [1.08, 1.38]	
9. Conclusions were based on 95% Bayesian credible intervals					0.95 [0.84, 1.08]	
8. Conclusions were based on Bayes's factors					0.74 [0.63, 0.84]	
62. Phenomenon is caused by single underlying factor					0.69 [0.57, 0.83]	
41. Study was published in high-impact-factor journal					−0.69 [−0.81, −0.58]	
7. Conclusions were based on effect size						1.09 [0.97, 1.2]
42. Study was published in low-impact-factor journal						1.02 [0.92, 1.13]
76. Replicating team was supportive of original result						0.90 [0.79, 1.01]
6. Conclusions were based on 95% confidence intervals						0.87 [0.76, 0.97]
51. Institution has a strong reputation for research						0.81 [0.72, 0.90]
70. Original result is "classic" finding						0.78 [0.66, 0.90]
33. Original study was published after 2010						0.68 [0.58, 0.77]
56. Theoretical justification for predicting more than one outcome						0.57 [0.45, 0.71]
28. Lead author was a senior researcher						0.56 [0.49, 0.65]
46. Study was from biological psychology						0.43 [0.33, 0.53]
23. Results have practical implications						0.42 [0.32, 0.53]
44. Study was from human factors psychology						0.28 [0.19, 0.36]
27. Lead author was an early-career researcher						0.25 [0.18, 0.32]
29. Original article was over 25 pages						0.23 [0.16, 0.30]
39. Lead author was right-handed						0.12 [0.07, 0.17]
40. Lead author was left-handed						

Note: 95% Bayesian credible intervals are in brackets.

Table 2. Factor Correlations With 95% Bayesian Credible Intervals [Lower, Upper]

Factor	1	2	3	4	5	6
1. Weak Methodology	—					
2. Potential for Questional Research Practices	.86 [.82, .88]					
3. Rigorous Analysis	−.86 [−.89, −83]	−.75 [−.79, −.70]				
4. Ease of Replication	−.91 [−.93, −.89]	−.66 [−.72, −.61]	.92 [.90, .94]			
5. Robustness of Conclusions	−.81 [−.84, −.77]	−.53 [−.60, −.46]	.88 [.85, .90]	.95 [.94, .96]		
6. Established Convention	−.42 [−.50, −.34]	−.04 [−.14, .05]	.62 [.56, .68]	.70 [.64, .75]	.82 [.78, .85]	-

Research Question 3: Are There Individual Differences in the Issues That People Consider When Evaluating Replicability?

To answer this research question, we conducted a cluster analysis using the Gaussian multivariate mixture modeling approach implemented via the *mclust* package (Scrucca et al., 2016) in R. The goal of this analysis was to partition participants into groups according to their factor scores such that participants in the same group held more similar views to each other than to participants in other groups. The analysis was conducted in two steps. In the first step, we fitted a series of alternative models that instantiated different numbers of clusters (between one and nine; all with unrestricted covariance matrices). The goal of this step was to identify the model that best characterized participants' factor scores. Model selection was conducted by comparing the Bayesian information criterion (BIC) across alternatives, and the model with the lowest BIC was selected.[2]

The results of the model selection are shown in Figure 2. The winning model (BIC = 21,214.17) assumed that there were three unique clusters of participants and that the clusters varied in size, shape, and orientation. The next best model (BIC = 21,273.64) assumed that there were two unique clusters that also varied in size, shape, and orientation.

In the second step, we used the winning model from Step 1 to classify participants into clusters. For the positioning of participants within each cluster with respect to the scores on each factor and the relationship between factor scores, see Figure 3. As can be seen, Cluster 1 (52% of participants; shown in blue) and Cluster 2 (42% of participants; shown in red) tend to agree on how the different factors would affect replicability. Both groups showed decreases in confidence in response to items associated with the Weak Methodology and the Potential for QRPs factors and increases in confidence in response to items associated with the Rigorous Analysis, Ease of Replication, and Robustness of Conclusions factors (with neither group showing much change in confidence in response to the items associated with the Established Convention factor). However, the changes in confidence among participants in Cluster 2 tended to be more profound than the changes among participants in Cluster 1. In other words, if items within a factor tended to increase or decrease confidence, confidence tended to increase/decrease more for participants in Cluster 2 compared with Cluster 1 (i.e., participants in Cluster 2 were the most sensitive). The participants in Cluster 3 (6% of participants; shown in green) deviated from the pattern demonstrated by Clusters 1 and 2. These participants had more variable views in general but demonstrated a systematic tendency to be less dissuaded by the items relating to the Potential for QRPs factor. Note that the mode of the Cluster 3 distribution is above zero for this factor—which suggests that participants within this cluster responded more positively to items associated with this factor—whereas the modes for the distributions associated with Clusters 1 and 2 were below zero.

Having established which participants belonged to each cluster, we next examined whether the clusters differed on key demographics variables. Table 3 shows the mean and standard deviation of each of the demographic items in the survey broken down by cluster. Participants in each cluster were quite similar across demographic items. Participants in Cluster 3 seemed to be the most concerned with replicability and were also the most likely to preregister, make a priori hypotheses, and run formal power analyses, whereas participants in Cluster 1 tended to have the lowest scores across these items. Furthermore, participants in Cluster 3 had nearly half as many citations on average compared with Clusters 1 and 2. However, participants in Cluster 3 had the equal highest h-index, which seems to indicate researchers across all three clusters tended to be similarly as prolific according to these metrics. Note that there was

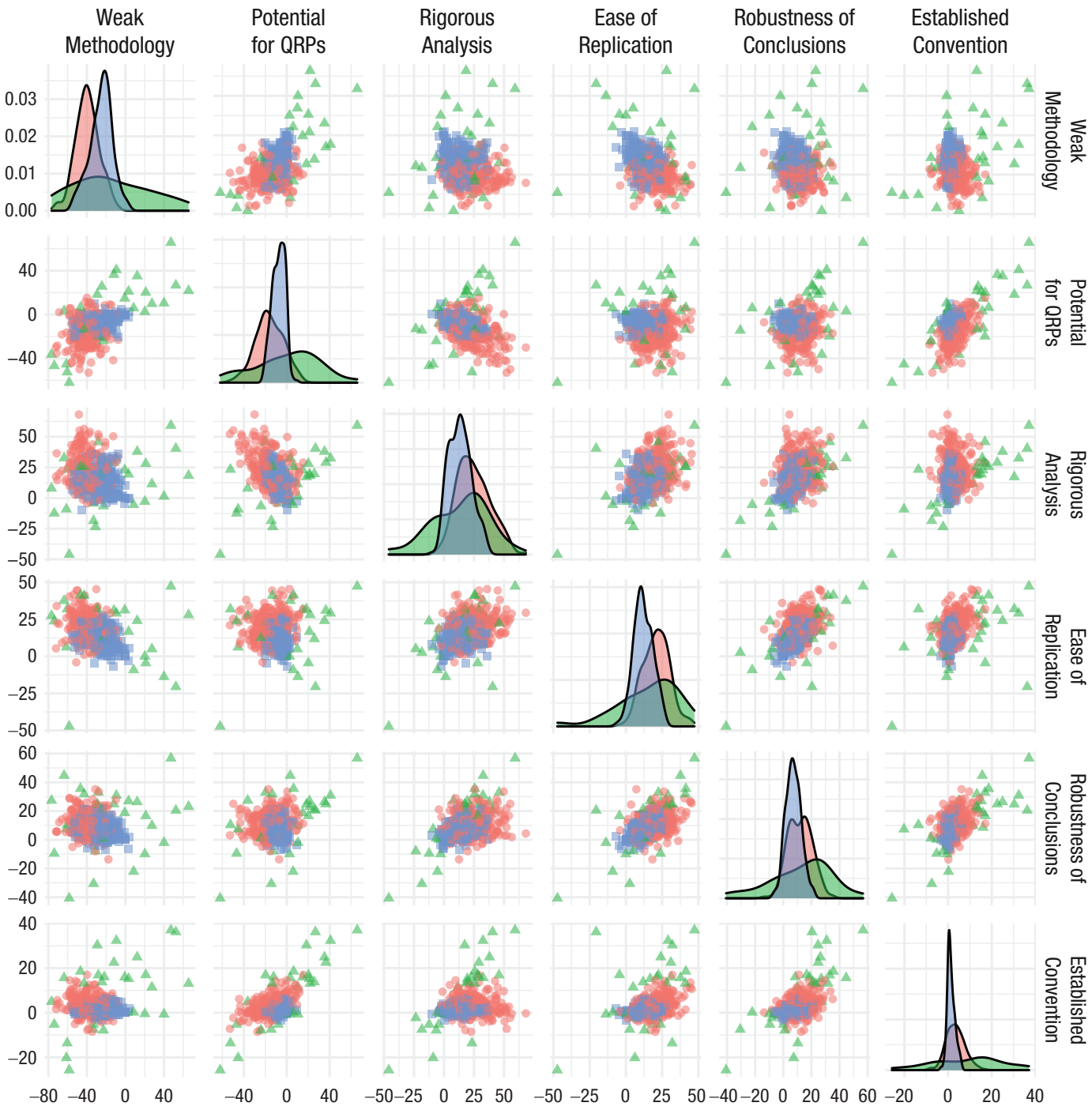

Fig. 3. Factor scores of participants in each cluster. The panels on the diagonal show the distribution of scores on the indicated factor. The other panels contain scatterplots showing the relationship between the factors. Plotted points representing participants in Clusters 1, 2, and 3 are shown in blue, red, and green, respectively.

a substantially smaller proportion of early-career and midcareer researchers in Cluster 3 but a higher proportion of postgraduate students and senior academics. Cluster 3 was also the smallest cluster, so it had the greatest vulnerability to random fluctuation. There did not seem to be any particularly substantial differences between clusters across subdisciplines.

Discussion

Accurate judgments of replicability are vital to ensure that researchers are building on work that is robust. Our aim was to understand how researchers make these judgments by assessing how 76 respective study attributes influence other psychological scientists' confidence in

Table 3. Differences in Demographic Variables Among the Different Clusters of Participants

Item	Cluster 1 *M* (*SD*)	Cluster 2 *M* (*SD*)	Cluster 3 *M* (*SD*)
Replication success of hypothetical psychological study	53.5 (18.1)	52.0 (18.1)	54.4 (25.8)
Age	47.0 (13.7)	44.3 (12.8)	47.3 (14.0)
Years since PhD	17.0 (14.0)	14.2 (13.5)	17.7 (13.0)
Familiarity with current debates surrounding replicability	6.6 (2.2)	7.0 (2.2)	6.4 (2.2)
Importance of replicability	8.0 (1.8)	8.6 (1.5)	8.9 (1.6)
Extent to which psychology is experiencing a replication crisis	6.3 (2.0)	7.0 (2.2)	7.3 (2.2)
Frequency of preregistering studies	3.5 (3.5)	4.0 (3.6)	5.3 (4.0)
Frequency of making a priori hypotheses	8.3 (1.9)	8.4 (2.0)	8.6 (1.4)
Frequency of running formal power analyses	6.3 (3.0)	6.3 (3.1)	8.0 (2.4)
h-index	24.5 (23.8)	22.3 (20.7)	24.5 (31.1)
Number of publications	68.3 (76.3)	66.3 (84.6)	65.1 (73.3)
Number of citations	5,058 (12,787)	4,927 (12,581)	2,740 (4,518)
Career stage (as % of cluster)			
Undergraduate			1%
Postgraduate	4%	5%	10%
Early-career academic	23%	30%	17%
Midcareer academic	29%	27%	21%
Senior academic	34%	28%	41%
Left academia before PhD		1%	
Left academia after PhD	3%	5%	3%
Psychology field (as % of cluster)			
Cognitive	18%	17%	17%
Clinical	15%	13%	14%
Developmental	6%	8%	7%
Industrial/organizational	10%	7%	14%
Evolutionary	5%	6%	
Neuropsychology	2%	2%	3%
Social	26%	22%	21%
Health	5%	9%	7%
Quantitative	1%	7%	7%
Human factors	1%	2%	7%
Biological	3%	3%	7%
Unsure	1%	1%	
Other	19%	2%	14%

the replicability of a finding. We focused on three research questions: How does each feature influence average confidence in replicability? What are the factors that influence judgments of replicability, and what are their constituent features? and Are there individual differences in the issues that people consider when evaluating replicability?

We found that there was considerable variability among the 76 study attributes in how they influenced researchers' confidence that a finding would replicate. Among the features that tended to result in the greatest increase in confidence was the existence of previous replications, high power or sample size, open data and materials, robustness of the phenomenon across contexts, a representative sample of participants, and analyses that were preregistered, planned in advance, or a Registered Report. Among the features that tended to result in the greatest decrease in confidence were the study having low power or a small sample size, data and methods that are not openly available, a phenomenon that is highly context dependent, or results that involve a three-way interaction, report inferences based on *p* values just below the threshold for significance, are surprising, or are based on post hoc analyses.

Furthermore, we found evidence for six somewhat distinct themes that psychological researchers consider when making judgments of replicability. The first theme was related to features that are often associated with weak methodology in the original study (e.g., low sample size/power, the use of a convenience sample) and

a lack of transparency (e.g., lack of open data/methods). The second theme consisted of features perceived as suggestive of QRPs such as HARKING or *p*-hacking (e.g., results that are surprising, involve complex interactions, are just below the threshold for significance, or are based on analyses that were not planned in advance). The third theme contained features that may have indicated a rigorous analysis (e.g., a large sample, high power, an analysis that was planned in advance). The fourth theme related to the ease of conducting a replication study (e.g., the existence of previous replications, the presence of open methods and data, and the effectiveness of the methodology). The fifth theme comprised study attributes that were indicative of the robustness of conclusions at a higher level (e.g., consistency with theory and prior research, robustness across contexts or experiments). The sixth and final theme, for which the underlying characteristics were hardest to identify, was mainly composed of features that might be associated with traditional markers of replicability (e.g., the status of the researcher or institution, the finding being regarded as classic, and the use of standard statistical tools such as effect size and confidence intervals).

We also found individual differences in how people were influenced by the criteria discussed above. Specifically, we identified three distinct types of responders who differed in the extent to which each theme influenced their confidence that a study would replicate. Participants in Cluster 1 (52.1%) and Cluster 2 (42.1%) responded in similar ways to each theme, but participants in Cluster 2 tended to have more extreme responses (i.e., their changes in confidence were more pronounced across different items). Participants in Cluster 3 (5.8%) held views that tended to be more variable but, on average, hardly changed across different attributes. Another noteworthy feature of participants in Cluster 3 was that they tended not to be dissuaded by items in the Potential for QRPs theme and even showed a slight average increase in confidence for those items. These results suggest that most of the psychological scientists in our sample tended to agree about the factors that influence replicability, although some seemed to be less dissuaded by what many would consider to be QRPs.

We found that the factors that most strongly influenced perceived replicability tended to relate to the factors that large-scale replication attempts have found to actually influence replicability. For example, a common theme found in these replication attempts was that a larger effect size, larger sample, and smaller *p* value in the original study were independently associated with a higher probability that the original study would successfully replicate (Camerer et al., 2018; Klein et al., 2014; Open Science Collaboration, 2015). Consistent with these findings, researchers in our sample tended to display an increase in confidence in the replicability of a finding if the *p* value was substantially below significance rather than "just" below significance and if the effect size or sample size was deemed as large instead of small. Furthermore, the Open Science Collaboration (2015) found that replication rates in social psychology tend to be lower than those in cognitive psychology, which seemed to be accurately reflected in the perceptions of the current study sample. Specifically, there was less confidence in the perceived replicability of a finding from social psychology compared with a finding from cognitive psychology. The Open Science Collaboration's replication attempt also tended to find that attributes relating to the experience of the authors was unrelated to the replicability of a result, which was also reflected in the perceptions of our sample.

The alignment of our sample's confidence with large-scale replication attempts suggests that psychological scientists are attuned to detecting research practices that are indicative of replicable research. This is supported by evidence from the aforementioned prediction market studies (Camerer et al., 2018; Dreber et al., 2015; Forsell et al., 2019) that have shown that psychological scientists are fairly accurate at predicting the outcomes of replication attempts. Indeed, our results may provide further insight into how much weight researchers give to certain features when making these judgments. Our findings shed light on study attributes that may be used to successfully predict whether a finding will replicate. However, further research is required to determine whether the study features that inspire the most or least confidence among researchers are the ones that are most successful in predicting, respectively, replication or failure to replicate.

Our results suggest that many of the practices that have been proposed as a means to improve the replicability of psychological research—such as open data and methods (Asendorpf et al., 2013; Miguel et al., 2014), preregistration and Registered Reports (Jonas & Cesario, 2016; Nosek et al., 2018; Wagenmakers et al., 2012), and basing conclusions on Bayesian inference (Etz & Vandekerckhove, 2016), or $p < .005$ rather than $p < .05$ (Benjamin et al., 2018)—do indeed improve confidence in replicability among our sample. Note, however, that the usefulness of some of these practices for improving confidence in psychological science has been disputed. For example, many authors have argued that preregistration is not as helpful as it is often purported and may actually be harmful insofar that it can instill misplaced confidence in research findings because it does little to promote improved theory development and does not necessarily ensure methodological rigor (MacEachern & Van Zandt, 2019; Szollosi & Donkin, 2021; Szollosi et al., 2020). This connects to a broader argument questioning

whether replicable results (i.e., effects or patterns in data) are even indicative of robust theory (i.e., explanatory frameworks for psychological phenomena) because (a) true psychological regularities can be, by nature, nonreplicable and (b) results that are actually false can be still be replicable (Devezer et al., 2020; Gervais, 2021). Hence, without an emphasis on rigorous theory, replication alone may not be a reliable indicator of robust research. Discussing the merits of these arguments is beyond the scope of this project, but we think that it is important to include these perspectives in conversations about replicability and the crisis of confidence, particularly if it comes to light that the practices people perceive to be effective at promoting replicability (or robust research more generally) do not align with practices that are actually effective.

Practical Implications

We found that a small subset of our participants (those in the "insensitive responders" cluster) were systematically less concerned with features that could indicate the potential for what many researchers might consider to be QRPs, such as *p*-hacking and HARKing. This could indicate that the replicability judgments of a substantial—albeit minority—group of psychological researchers might not be as accurate as they could be. This could also be taken as evidence that more needs to be done to communicate the harmful consequences of these practices to the broader scientific community. However, it is important to consider that whether a practice is deemed a QRP might be dependent on the subdiscipline or type of research that is being conducted. For example, in quantitative fields, exploration and post hoc adjustments are often necessary for model development and are often synonymous with best practice when done rigorously (MacEachern & Van Zandt, 2019; Navarro, 2020; Szollosi & Donkin, 2021). Therefore, it would be useful to gain a better understanding of which types of researchers are more tolerant of these so-called QRPs and why so we can have a more open and inclusive discussion about how to improve practices as a discipline.

Future Research

A useful next step might be to use these data to create a scale for measuring the different facets that determine perceived replicability. This would involve optimizing the item set so that the various facets can be reliably measured using fewer items. Future work might also benefit from examining perceptions of the replicability of more specific categories of studies, including studies from different subdisciplines. Our survey was agnostic with regard to characteristics of the study being replicated, such as the sample size, the design, the nature of the hypotheses, and the methodological approach. We designed the survey this way so our results would not be constrained to a specific type of study. However, it is possible that our use of a general study description, as opposed to a more detailed one, may mean that our results do not generalize to certain types of study designs or even to any specific study. It is, therefore, important for future work to examine how people make judgments regarding the replicability of results obtained using more specific research protocols.

Future research might also address the question of whether there are individual differences in the factor structure itself. Our analysis showed individual differences in the combinations of factors scores. However, our analysis assumed that all researchers rely on the same factor structure. Although we believe that this was an appropriate first step, it may be the case that there is variability across individuals in the factor structure itself. This variability can be examined in future research by conducting, for example, a mixture exploratory factor analysis, which allows for differences in the factor structure that emerges across individuals. This could also provide further clarity as to whether there are any differences in the perceived utility or harm of certain research practices across different subdisciplines. Indeed, our recruitment of such a broad sample may have resulted in a blending of factor structures that might truly differ across subsets of researchers.

It will also be important for future work to address potential moderating factors. For example, one might not care about the statistical approach if the sample size is sufficiently high but might care a great deal if the sample size is lower. It is also possible that familiarity with certain practices influences their perceived efficacy. For example, people more familiar with Registered Reports may be more likely to experience an increase in their confidence in a study's replicability upon finding out that the study was a Registered Report. We believe that an examination of moderating factors is a useful next step for future research.

Conclusion

The aim of this project was to understand how 76 different study attributes influenced psychological scientists' confidence that a finding would replicate. We found that, on average, these perceptions tended to match what large-scale replication attempts have found to actually influence replicability. Our sample also showed confidence in open science practices like openly available data and methods, preregistration, and Registered Reports. Overall, we found evidence for six themes that psychological scientists consider when evaluating

replicability and that there may be individual differences in how these themes influence subsets of researchers. This work provides a useful starting point for understanding how people judge the replicability of a research finding. We hope that by better understanding researchers' confidence in these practices, this will inform what still needs to be done to strengthen faith in practices that improve replicability but also reduce confidence in practices that might not be robust.

Appendix A: Survey Items

The full list of items that were used to assess the effect of study attributes on confidence in replicability is presented in Table A1. The items were responded to on an 11-point scale ranging from −5 (*substantial decrease in confidence*) to 5 (*substantial increase in confidence*); 0 represented *no change in confidence*. The survey items were presented in random order.

Appendix B: Priors for the Exploratory Factor Analysis

We used the default prior settings within the *BayesFM package* for the exploratory factor analysis. Table B1 contains a description of the parameters in the model used for the exploratory factor analysis and the prior on each parameter.

Table A1. Items Used to Assess Effect of Study Attributes on Confidence in Replicability

Item
1. The original study was pre-registered.
2. The original study was not pre-registered.
3. The original study was a Registered Report.
4. The original study was not a Registered Report.
5. The data from the original study was analyzed using null-hypothesis testing methods and conclusions were based on an interpretation of *p*-values being < .05.
6. The data from the original study was analyzed using null-hypothesis testing methods and conclusions were based on an interpretation of 95% confidence intervals.
7. The data from the original study was analyzed using null-hypothesis testing methods and analyses included measures of effect size (e.g., partial eta squared, or *r*-squared).
8. The data from the original study were analyzed using Bayesian methods and conclusions were based on Bayes factors being greater than 3.
9. The data from the original study were analyzed using Bayesian methods and conclusions were based on Bayesian 95% credible intervals.
10. The conclusions of the original study were based on analyses planned before data collection.
11. The conclusions of the original study were based on post-hoc analyses.
12. The original study was from the field of cognitive psychology.
13. The original study was from the field of social psychology.
14. The original study was from the field of neuroscience or neuropsychology.
15. The original study was from the field of developmental psychology.
16. The original study was from the field of evolutionary psychology.
17. The original study was from the field of clinical psychology.
18. The original study was from the field of industrial/organizational psychology.
19. The main result reported by the original study was highly surprising.
20. The original study reported a result that was consistent with (or predicted by) existing theory.
21. The original study had low statistical power.
22. The original study had high statistical power.
23. The results of the original study have direct or immediate practical applications (e.g., improving mental health).
24. The primary result reported in the original study was a main effect (i.e., not an interaction).
25. The primary result reported in the original study was a two-way interaction.
26. The primary result reported in the original study was a three-way interaction.
27. The lead author on the original study was an early career researcher (with 0-5 years of research experience post PhD).
28. The lead author on the original study was a senior researcher (with at least 15 years of research experience post PhD).
29. The original study was published in a journal article that was over 25 pages in length.
30. The original study was published in a journal article that was less than 8 pages in length.
31. The original study was published before 2000.

(continued)

Table A1. *(continued)*

Item
32. The original study was published between 2000 and 2010.
33. The original study was published after 2010.
34. The data from the original study was made available.
35. The data from the original study was not made available.
36. The full methods from the original study were made available.
37. The full methods from the original study were not made available.
38. The data from the original study was analyzed using null-hypothesis testing methods and conclusions were based on an interpretation of *p*-values being < .005 (instead of $p < .05$).
39. The lead author on the original study was right handed.
40. The lead author on the original study was left handed.
41. The original study was published in a journal with a high impact factor (relative to discipline).
42. The original study was published in a journal with a low impact factor (relative to discipline).
43. The original study was from the field of health psychology.
44. The original study was from the field of human factors psychology.
45. The original study was from the field of quantitative psychology.
46. The original study was from the field of biological psychology.
47. In the original study the *p*-value was just below the threshold of significance.
48. In the original study the *p*-value was far below the threshold of significance.
49. The original study had a large sample size.
50. The original study has a small sample size.
51. The lead author of the original study was based at an institution with a strong reputation for research.
52. The lead author of the original study was based at an institution without a strong reputation for research.
53. The original study used a convenience sample of undergraduate psychology participants.
54. The original study used an online convenience sample recruited via Mechanical Turk.
55. The original study used a recruitment procedure that ensured the sample was representative of the target population.
56. In the original study there was a theoretical justification for predicting more than one pattern of outcomes (i.e., competing predictions).
57. The original study used commonly used manipulations and measures.
58. The original study reported evidence that the manipulations and measures effectively manipulated or assessed the intended constructs.
59. The original study reported pilot evidence that the procedure and measures were operating as intended.
60. The original study design explicitly built upon prior research.
61. The original study design was creative/novel.
62. The observed phenomenon is believed to be caused by a single underlying factor.
63. The observed phenomenon is believed to be caused by multiple underlying factors.
64. The observed phenomenon is believed to be highly context dependent.
65. The observed phenomenon is believed to be robust across contexts.
66. The original study involved multiple experiments.
67. The original study involved a single experiment.
68. The original result received considerable media attention.
69. The original study was conducted by a collaborator or close colleague.
70. The original result is regarded as a "classic" finding.
71. The original result has been successfully and independently replicated using methods that were as close to the original study as possible (i.e., a direct replication).
72. The original result has been successfully and independently replicated in a novel context (i.e., a conceptual replication).
73. The original result had a relatively small effect size.
74. The original result had a relatively large effect size.
75. The researchers conducting the replication are known to be skeptical of the original result.
76. The researchers conducting the replication are known to be supportive of the original result.

Table B1. Description of Parameters in the Exploratory Factor Model

Parameter	Prior Value
Scaling parameters of the variance of the Normal prior on the nonzero factor loadings	10
Shape parameters of the Inverse-Gamma prior on the idiosyncratic variances	2
Scale parameters of the Inverse-Gamma prior on the idiosyncratic variances	1
Degrees of freedom of the Inverse-Wishart prior on the covariance matrix of the latent factors in the expanded model	7
Scale parameters of the Inverse-Wishart prior on the covariance matrix of latent factors in the expanded model	1
First shape parameter of the Beta prior distribution on the probability that a manifest variable does not load on any factor	2
Second shape parameter of the Beta prior distribution on the probability that a manifest variable does not load on any factor	1
Concentration parameters of the Dirichlet prior distribution on the indicators	1/6

Transparency

Action Editor: Daniel J. Simons
Editor: Daniel J. Simons

Author Contributions

Stage 1: M. Alister and R. Vickers-Jones contributed equally to the stage 1 registered report. All authors contributed to the development of the survey. R. Vickers-Jones drafted the introduction, M. Alister drafted the method section and "demographic and background information" section of the expected results, all of which were edited by T. Ballard and D. K. Sewell. M. Alister extracted participant emails from the Web of Science, R. Vickers-Jones tracked email response rates. T. Ballard wrote the script for figure 1, and the EFA and cluster analysis scripts, and drafted the corresponding sections in the expected results, with consultation from D. K. Sewell. T. Ballard and D. K. Sewell co-drafted the stage-1 discussion which was edited by M. Alister and R. Vickers-Jones. Stage 2: M. Alister and T. Ballard performed the analyses for the demographic and background information, and part 1 of the results section. T. Ballard performed the part 2 and 3 analyses. M. Alister drafted the stage 2 discussion, which was edited by T. Ballard and D. K. Sewell. All of the authors approved the final manuscript for submission.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

T. Ballard was supported by a Discovery Early Career Researcher Award from the Australian Research Council (DE180101340).

Open Practices

Open Data: https://osf.io/g9rsc
Open Materials: https://osf.io/g9rsc
Preregistration: https://osf.io/g9rsc

All data have been made publicly available via OSF and can be accessed at https://osf.io/g9rsc. All materials have been made publicly available via OSF and can be accessed at https://osf.io/g9rsc. The analysis plan was preregistered at OSF prior to data collection and can be accessed at https://osf.io/g9rsc. This article has received badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iDs

Manikya Alister https://orcid.org/0000-0001-8630-9938
Raine Vickers-Jones https://orcid.org/0000-0003-4492-6123
Timothy Ballard https://orcid.org/0000-0001-8875-4541

Acknowledgments

We thank Jessica Mead for her assistance in constructing the initial version of the survey. We also thank the repliCATS team, Norbert Schwarz, Simine Vazire, Jeff Sherman, Roger Giner-Sorolla, Sanjay Srivastava, Jason Tangen, and Brian Nosek, who contributed items based on their various areas of expertise.

Notes

1. We interpreted this factor as a Questionable Research Practices factor because, on average, these items tended to show a decrease in confidence for our sample. However, we acknowledge that none of these things, by themselves, imply any wrongdoing on the part of a researcher, and for many people, they might not necessarily constitute questionable research practices.
2. We conducted a model recovery analysis to determine the precision with which the number of clusters could be measured given the number of participants in our sample and the number of factors that emerged from the previous analysis. We simulated 100 synthetic data sets with one, two, three, or four clusters of participants and fitted the clustering model to each data set (for a total of 400 data sets). The true number of clusters was recovered 97% of the time. This result suggests that the number of clusters should be able to be reliably estimated in our data set.

References

Asendorpf, J. B., Conner, M., Fruyt, F. D., Houwer, J. D., Denissen, J. J. A., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., Aken, M. A. G. van, Weber, H., & Wicherts,

J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108–119. https://doi.org/10.1002/per.1919

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., Boeck, P. D., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., . . . Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., . . . Wu, H. (2018). Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145–153. https://doi.org/10.1037/h0045186

Conti, G., Frühwirth-Schnatter, S., Heckman, J. J., & Piatek, R. (2014). Bayesian exploratory factor analysis. *Journal of Econometrics*, *183*(1), 31–57. https://doi.org/10.1016/j.jeconom.2014.06.008

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Buzbas, E. O. (2020). *The case for formal methodology in scientific reform*. BioRxiv. https://doi.org/10.1101/2020.04.26.048306

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B. A., & Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences, USA*, *112*(50), Article 15343. https://doi.org/10.1073/pnas.1516179112

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*, Article 621. https://doi.org/10.3389/fpsyg.2015.00621

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., . . . Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the Reproducibility Project: Psychology. *PLOS ONE*, *11*(2), Article e0149794. https://doi.org/10.1371/journal.pone.0149794

Field, S. M., Wagenmakers, E.-J., Kiers, H., Hoekstra, R., Ernst, A., & van Ravenzwaaij, D. (2018). *The effect of preregistration on trust in empirical research findings*. OSF. https://doi.org/10.17605/OSF.IO/B3K75

Forsell, E., Viganola, D., Pfeiffer, T., Almenberg, J., Wilson, B., Chen, Y., Nosek, B. A., Johannesson, M., & Dreber, A. (2019). Predicting replication outcomes in the Many Labs 2 study. *Replications in Economic Psychology and Behavioral Economics*, *75*, Article 102117. https://doi.org/10.1016/j.joep.2018.10.009

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Gervais, W. M. (2021). Practical methodological reform needs good theory. *Perspectives on Psychological Science*. Advance online publication. https://doi.org/10.1177/1745691620977471

Jonas, K. J., & Cesario, J. (2016). How can preregistration contribute to research in our field? *Comprehensive Results in Social Psychology*, *1*(1–3), 1–7. https://doi.org/10.1080/23743603.2015.1070611

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., . . . Nosek, B. A. (2014). Investigating variation in replicability: A "Many Labs" replication project. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., . . . Nosek, B. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Academic Press.

MacEachern, S. N., & Van Zandt, T. (2019). Preregistration of modeling exercises may not be useful. *Computational Brain & Behavior*, *2*(3), 179–182. https://doi.org/10.1007/s42113-019-00038-x

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D. P., Humphreys, M., Imbens, G., Laitin, D., Madon, T., Nelson, L., Nosek, B. A., Petersen, M., Sedlmayr, R., Simmons, J. P., Simonsohn, U., & Laan, M. V. der. (2014). Promoting transparency in social science research. *Science*, *343*(6166), 30–31. https://doi.org/10.1126/science.1245317

National Science Foundation. (2017). *Survey of doctorate recipients*. https://nsf.gov/statistics/srvydoctoratework/#sd

Navarro, D. (2020). *If mathematical psychology did not exist we might need to invent it: A comment on theory building in psychology*. PsyArXiv. https://doi.org/10.31234/osf.io/ygbjp

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the*

National Academy of Sciences, USA, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), Article aac4716. https://doi.org/10.1126/science.aac4716

Pashler, H., & Wagenmakers, E. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Piatek, R. (2019). *BayesFM: Bayesian inference for factor modeling* (R package Version 0.1.3) [Computer software]. https://cran.r-project.org/web/packages/BayesFM/index.html

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, *8*(1), 289–317.

Szollosi, A., & Donkin, C. (2021). Arrested theory development: The misguided distinction between exploratory and confirmatory research. *Perspectives on Psychological Science*. Advance online publication. https://doi.org/10.1177/1745691620966796

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2020). Is preregistration worthwhile? *Trends in Cognitive Sciences*, *24*(2), 94–95. https://doi.org/10.1016/j.tics.2019.11.009

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105–110. https://doi.org/10.1037/h0031322

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638. https://doi.org/10.1177/1745691612463078